



# UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE  
United States Patent and Trademark Office  
Address: COMMISSIONER FOR PATENTS  
P.O. Box 1450  
Alexandria, Virginia 22313-1450  
[www.uspto.gov](http://www.uspto.gov)

b

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
10/814,960	03/31/2004	Robert Boman	9432-000270	8323
27572	7590	01/22/2008	EXAMINER	
HARNESS, DICKEY & PIERCE, P.L.C.			COLUCCI, MICHAEL C	
P.O. BOX 828			ART UNIT	PAPER NUMBER
BLOOMFIELD HILLS, MI 48303			2626	
MAIL DATE		DELIVERY MODE		
01/22/2008		PAPER		

**Please find below and/or attached an Office communication concerning this application or proceeding.**

The time period for reply, if any, is set in the attached communication.

<b>Office Action Summary</b>	Application No.	Applicant(s)
	10/814,960	BOMAN ET AL.
	Examiner Michael C. Colucci	.Art Unit 2626

-- The MAILING DATE of this communication appears on the cover sheet with the correspondence address --

#### Period for Reply

A SHORTENED STATUTORY PERIOD FOR REPLY IS SET TO EXPIRE 3 MONTH(S) OR THIRTY (30) DAYS, WHICHEVER IS LONGER, FROM THE MAILING DATE OF THIS COMMUNICATION.

- Extensions of time may be available under the provisions of 37 CFR 1.136(a). In no event, however, may a reply be timely filed after SIX (6) MONTHS from the mailing date of this communication.
- If NO period for reply is specified above, the maximum statutory period will apply and will expire SIX (6) MONTHS from the mailing date of this communication.
- Failure to reply within the set or extended period for reply will, by statute, cause the application to become ABANDONED (35 U.S.C. § 133). Any reply received by the Office later than three months after the mailing date of this communication, even if timely filed, may reduce any earned patent term adjustment. See 37 CFR 1.704(b).

#### Status

- 1) Responsive to communication(s) filed on \_\_\_\_\_.
- 2a) This action is FINAL.                    2b) This action is non-final.
- 3) Since this application is in condition for allowance except for formal matters, prosecution as to the merits is closed in accordance with the practice under *Ex parte Quayle*, 1935 C.D. 11, 453 O.G. 213.

#### Disposition of Claims

- 4) Claim(s) 1-27 is/are pending in the application.
- 4a) Of the above claim(s) \_\_\_\_\_ is/are withdrawn from consideration.
- 5) Claim(s) \_\_\_\_\_ is/are allowed.
- 6) Claim(s) 1-27 is/are rejected.
- 7) Claim(s) \_\_\_\_\_ is/are objected to.
- 8) Claim(s) \_\_\_\_\_ are subject to restriction and/or election requirement.

#### Application Papers

- 9) The specification is objected to by the Examiner.
- 10) The drawing(s) filed on 31 March 2004 is/are: a) accepted or b) objected to by the Examiner.  
Applicant may not request that any objection to the drawing(s) be held in abeyance. See 37 CFR 1.85(a).  
Replacement drawing sheet(s) including the correction is required if the drawing(s) is objected to. See 37 CFR 1.121(d).
- 11) The oath or declaration is objected to by the Examiner. Note the attached Office Action or form PTO-152.

#### Priority under 35 U.S.C. § 119

- 12) Acknowledgment is made of a claim for foreign priority under 35 U.S.C. § 119(a)-(d) or (f).  
a) All    b) Some \* c) None of:  
  1. Certified copies of the priority documents have been received.
  2. Certified copies of the priority documents have been received in Application No. \_\_\_\_\_.
  3. Copies of the certified copies of the priority documents have been received in this National Stage application from the International Bureau (PCT Rule 17.2(a)).

\* See the attached detailed Office action for a list of the certified copies not received.

#### Attachment(s)

- |   |   |
|---|---|
| 1) <input checked="" type="checkbox"/> Notice of References Cited (PTO-892)   | 4) <input type="checkbox"/> Interview Summary (PTO-413)<br>Paper No(s)/Mail Date: _____ |
| 2) <input type="checkbox"/> Notice of Draftsperson's Patent Drawing Review (PTO-948)  | 5) <input type="checkbox"/> Notice of Informal Patent Application                       |
| 3) <input checked="" type="checkbox"/> Information Disclosure Statement(s) (PTO/SB/08)<br>Paper No(s)/Mail Date <u>03/31/2004</u> . | 6) <input type="checkbox"/> Other: _____  |

## DETAILED ACTION

### ***Claim Rejections - 35 USC § 102***

1. The following is a quotation of the appropriate paragraphs of 35 U.S.C. 102 that form the basis for the rejections under this section made in this Office action:

A person shall be entitled to a patent unless –

(b) the invention was patented or described in a printed publication in this or a foreign country or in public use or on sale in this country, more than one year prior to the date of application for patent in the United States.

2. Claims 1 and 21-23 rejected under 35 U.S.C. 102(b) as being anticipated by Van Thong US 6490553 B2 (hereinafter Van Thong).

Re claim 1, Van Thong teaches a media production system, comprising:

a textual alignment module aligning multiple speech recordings to textual lines of a script based on speech recognition results (col 3 lines 25-31 & Fig. 1 items 13, 15, and 17);

a navigation module responding to user navigation selections respective of the textual lines of the script by communicating to the user corresponding (col 8 lines 18-31), line-specific portions of the multiple speech recordings (col 7 lines 46-60);

an editing module responding to user associations (col 8 lines 18-31) of multiple speech recordings with textual lines by accumulating line-specific portions of the multiple speech recordings (col 7 lines 46-60) in a combination recording based on at least one of relationships of textual lines in the script to the combination recording (col 3 lines 25-31 & Fig. 1 items 13, 15, and 17), and temporal alignments between the multiple speech recordings and the combination recording (col 1 lines 20-46).

Re claim 21, Van Thong teaches the system of claim 1, wherein said navigation module is adapted to play a user-specified portion (col 8 lines 18-31) of a speech recording in response to a sample request (col 18 lines 31-53).

Re claim 22, Van Thong teaches the system of claim 1, wherein said navigation module is adapted to play at least one of a user-specified section of the combination recording and a preview (col 1 lines 20-30) of the user-specified section based on a sequence of portions of multiple speech recordings (col 5 line 34-62).

Re claim 23, Van Thong teaches the system of claim 1, wherein said navigation module is adapted to record final selection of at least one of a speech recording and a specific portion thereof with respect to a textual line. (col 10 lines 19-47).

***Claim Rejections - 35 USC § 103***

3. The following is a quotation of 35 U.S.C. 103(a) which forms the basis for all obviousness rejections set forth in this Office action:

(a) A patent may not be obtained though the invention is not identically disclosed or described as set forth in section 102 of this title, if the differences between the subject matter sought to be patented and the prior art are such that the subject matter as a whole would have been obvious at the time the invention was made to a person having ordinary skill in the art to which said subject matter pertains. Patentability shall not be negatived by the manner in which the invention was made.

4. Claims 2, 3, 9, 10, 17, and 20 rejected under 35 U.S.C. 103(a) as being unpatentable over Van Thong US 6490553 B2 (hereinafter Van Thong) in view of Perez-Mendez et al US 5754978 A (hereinafter Perez).

Re claims 2, 9, and 20, Van Thong fails to teach the system of claim 1, further comprising a ranking module adapted to tag at least one of speech recordings and specific portions thereof with ranking data. (Perez col 10 line 61 - col 11 line 11).

NOTE: For purposes of prior art, a navigation module is construed to both functionally equivalent and equally effective as the system in (Van Thong Fig. 1), where a user and system respond in accordance with commands and/or text.

Perez teaches that the "n" top choices refers to a variable number of possibilities, which can be supplied by a speech recognition engine tagged with a probability or in a ranked order of descending probability of being correct. For example, the comparison might require that the top two choices from each engine match before accepting the first as the accepted and recognized text. Alternatively, if the top choice of each do not match, it might be acceptable if one of the engines has a second choice that matches the other engine's top choice.

Therefore, it would have been obvious to one of ordinary skill in the art at the time of the invention a ranking module that tags speech recordings with ranking data. Tagging ranked data would allow for recognition of the best choices of speech recordings and allows for book marking or the use of keywords or key terms as a form of indicating the most acceptable data in a summarized manner.

Re claim 3, Van Thong fails to teach the system of claim 2, wherein said ranking module is adapted to recognize tags (Perez col 2 lines 1-14) associated with the speech recordings and tag at least one of speech recordings and specific portions thereof accordingly.

Perez teaches that the "n" top choices refers to a variable number of possibilities, which can be supplied by a speech recognition engine tagged with a probability or in a ranked order of descending probability of being correct. For example, the comparison might require that the top two choices from each engine match before accepting the first as the accepted and recognized text. Alternatively, if the top choice of each do not match, it might be acceptable if one of the engines has a second choice that matches the other engine's top choice.

Additionally Perez teaches probability determined during the decoding can be used as a score for the match between the input speech and the chosen sentence text.

NOTE: For purposes of prior art, recognizing a tagged item is construed to be both functionally equivalent and equally effective as ranking a set of data that is tagged with a probability, where the probability is tagged with a score. In order to rank a set of data, a probability is assigned, where recognition of tags would be necessary in order to achieve a proper ranking scheme.

Therefore, it would have been obvious to one of ordinary skill in the art at the time of the invention a ranking module that tags speech recordings with ranking data. Tagging ranked data would allow for recognition of the best choices of speech

recordings and allows for book marking or the use of keywords or key terms as a form of indicating the most acceptable data in a summarized manner.

Re claim 10, Van Thong teaches the system of claim 9, further wherein said navigation module is adapted to rank at least one of speech, recordings and specific portions thereof based on order in which the speech recordings were produced (col 5 lines 12-31).

Re claim 17, Van Thong teaches speech recording production personnel during a speech recording process (col 8 lines 18-31).

However, Van Thong fails to teach the system of claim 9, wherein said navigation module is adapted to rank at least one of speech recordings and specific portions thereof based on ranking tags supplied thereto (Perez col 2 lines 1-14).

Perez teaches that the "n" top choices refers to a variable number of possibilities, which can be supplied by a speech recognition engine tagged with a probability or in a ranked order of descending probability of being correct. For example, the comparison might require that the top two choices from each engine match before accepting the first as the accepted and recognized text. Alternatively, if the top choice of each do not match, it might be acceptable if one of the engines has a second choice that matches the other engine's top choice.

Therefore, it would have been obvious to one of ordinary skill in the art at the time of the invention a navigation module that ranks speech recordings based on

ranked tags supplied to personnel. Tagging ranked data would allow for recognition of the best choices of speech recordings and allows for book marking by a user, or the use of keywords or key terms as a form of indicating the most acceptable data in a summarized manner for a user.

**5. Claims 4-8, 11-16, and 18 rejected under 35 U.S.C. 103(a) as being unpatentable over Van Thong US 6490553 B2 (hereinafter Van Thong) in view of Perez-Mendez et al US 5754978 A (hereinafter Perez) and further in view of Bakis US 6556972 B1 (hereinafter Bakis).**

Re claim 4, Van Thong fails to teach the system of claim 3, wherein said ranking module is adapted to recognize voice tags (Perez col 2 lines 1-14).

Perez teaches that the "n" top choices refers to a variable number of possibilities, which can be supplied by a speech recognition engine tagged with a probability or in a ranked order of descending probability of being correct. For example, the comparison might require that the top two choices from each engine match before accepting the first as the accepted and recognized text. Alternatively, if the top choice of each do not match, it might be acceptable if one of the engines has a second choice that matches the other engine's top choice.

Additionally Perez teaches probability determined during the decoding can be used as a score for the match between the input speech and the chosen sentence text.

NOTE: For purposes of prior art, recognizing a tag is construed to be both functionally equivalent and equally effective as ranking a set of data that is tagged with

a probability, where the probability is tagged with a score. In order to rank a set of data, a probability is assigned, where recognition of tags would be necessary in order to achieve a proper ranking scheme.

However, Van Thong in view of Perez fails to teach voice tags based on key phrases (Bakis col 4 lines 41-64).

NOTE: For purposes of prior art, a key phrase is construed to both functionally equivalent and equally effective as a phrase containing unique information relating to pitch, duration, speech rate, or mood/emotion.

Bakis teaches multiple output sentences for a given word or phrase, where each output sentence for a given word or phrase reflects a different emotional emphasis and could be selected automatically, or manually as desired, to create a specific emotional effect. For example, the same output sentence for a given word or phrase can be recorded three times, to selectively reflect excitement, sadness or fear. In further variations, the same output sentence for a given word or phrase can be recorded to reflect different accents, dialects, pitch, loudness or rates of speech. Changes in the volume or pitch of speech can be utilized, for example, to indicate a change in the importance of the content of the speech. The variable rate of speech outputs can be used to select a translation that has a best fit with the spoken phrase.

Therefore, it would have been obvious to one of ordinary skill in the art at the time of the invention a ranking module adapted to recognize voice tags based on key phrases. Tagging ranked data would allow for recognition of the best choices of speech

recordings and allows for book marking or the use of keywords or key terms as a form of indicating the most acceptable data in a summarized manner. Additionally, recognizing key phrases would allow for the omission of less important words during the ranking of a speech data set.

Re claim 5, Van Thong fails to teach the system of claim 2, wherein said ranking module is adapted to recognize key phrases within the speech recordings and tag at least one of speech recordings and specific portions thereof accordingly (Perez col 2 lines 1-14).

Perez teaches that the "n" top choices refers to a variable number of possibilities, which can be supplied by a speech recognition engine tagged with a probability or in a ranked order of descending probability of being correct. For example, the comparison might require that the top two choices from each engine match before accepting the first as the accepted and recognized text. Alternatively, if the top choice of each do not match, it might be acceptable if one of the engines has a second choice that matches the other engine's top choice.

Additionally Perez teaches probability determined during the decoding can be used as a score for the match between the input speech and the chosen sentence text.

NOTE: For purposes of prior art, recognizing a tag is construed to be both functionally equivalent and equally effective as ranking a set of data that is tagged with a probability, where the probability is tagged with a score. In order to rank a set of data,

a probability is assigned, where recognition of tags would be necessary in order to achieve a proper ranking scheme.

However, Van Thong in view of Perez fails to teach a module is adapted to recognize key phrases (Bakis col 4 lines 41-64).

NOTE: For purposes of prior art, recognizing a tagged item is construed to be both functionally equivalent and equally effective as ranking a set of data that is tagged with a probability, where the probability is tagged with a score. In order to rank a set of data, a probability is assigned, where recognition of tags would be necessary in order to achieve a proper ranking scheme.

Bakis teaches multiple output sentences for a given word or phrase, where each output sentence for a given word or phrase reflects a different emotional emphasis and could be selected automatically, or manually as desired, to create a specific emotional effect. For example, the same output sentence for a given word or phrase can be recorded three times, to selectively reflect excitement, sadness or fear. In further variations, the same output sentence for a given word or phrase can be recorded to reflect different accents, dialects, pitch, loudness or rates of speech. Changes in the volume or pitch of speech can be utilized, for example, to indicate a change in the importance of the content of the speech. The variable rate of speech outputs can be used to select a translation that has a best fit with the spoken phrase.

Therefore, it would have been obvious to one of ordinary skill in the art at the time of the invention a ranking module adapted to recognize key phrases. Tagging

ranked data would allow for recognition of the best choices of speech recordings and allows for book marking or the use of keywords or key terms as a form of indicating the most acceptable data in a summarized manner. Additionally, recognizing key phrases would allow for the omission of less important words during the ranking of a speech data set.

Re claims 6 and 12, Van Thong fails to teach a ranking module (Perez col 2 lines 1-14).

Perez teaches that the "n" top choices refers to a variable number of possibilities, which can be supplied by a speech recognition engine tagged with a probability or in a ranked order of descending probability of being correct. For example, the comparison might require that the top two choices from each engine match before accepting the first as the accepted and recognized text. Alternatively, if the top choice of each do not match, it might be acceptable if one of the engines has a second choice that matches the other engine's top choice.

Additionally Perez teaches probability determined during the decoding can be used as a score for the match between the input speech and the chosen sentence text.

NOTE: For purposes of prior art, recognizing a tag is construed to be both functionally equivalent and equally effective as ranking a set of data that is tagged with a probability, where the probability is tagged with a score. In order to rank a set of data,

a probability is assigned, where recognition of tags would be necessary in order to achieve a proper ranking scheme.

However, Van Thong in view of Perez fails to teach the system of claim 2, wherein said ranking module is adapted to evaluate pitch of speech within the speech recordings and tag at least one of speech recordings and specific portions thereof accordingly (Bakis col 4 lines 41-64).

Bakis teaches multiple output sentences for a given word or phrase, where each output sentence for a given word or phrase reflects a different emotional emphasis and could be selected automatically, or manually as desired, to create a specific emotional effect. For example, the same output sentence for a given word or phrase can be recorded three times, to selectively reflect excitement, sadness or fear. In further variations, the same output sentence for a given word or phrase can be recorded to reflect different accents, dialects, pitch, loudness or rates of speech. Changes in the volume or pitch of speech can be utilized, for example, to indicate a change in the importance of the content of the speech. The variable rate of speech outputs can be used to select a translation that has a best fit with the spoken phrase.

Therefore, it would have been obvious to one of ordinary skill in the art at the time of the invention a ranking module adapted to evaluate pitch of speech and tag an according speech portion. Tagging ranked data would allow for recognition of the best choices of speech recordings and allows for book marking or the use of keywords or key terms as a form of indicating the most acceptable data in a summarized manner.

Additionally, recognizing various prosodic features such as pitch, intonation, emotion, stress, speech rate, duration, etc. would allow for the omission of less important words during the ranking of a speech data set, where a tag can be applied more precisely prior to ranking (i.e. finding a phrase with specific pitch level, duration, and emotion).

Re claims 7 and 13, Van Thong fails to teach a ranking module (Perez col 2 lines 1-14).

Perez teaches that the "n" top choices refers to a variable number of possibilities, which can be supplied by a speech recognition engine tagged with a probability or in a ranked order of descending probability of being correct. For example, the comparison might require that the top two choices from each engine match before accepting the first as the accepted and recognized text. Alternatively, if the top choice of each do not match, it might be acceptable if one of the engines has a second choice that matches the other engine's top choice.

Additionally Perez teaches probability determined during the decoding can be used as a score for the match between the input speech and the chosen sentence text.

NOTE: For purposes of prior art, recognizing a tag is construed to be both functionally equivalent and equally effective as ranking a set of data that is tagged with a probability, where the probability is tagged with a score. In order to rank a set of data, a probability is assigned, where recognition of tags would be necessary in order to achieve a proper ranking scheme.

However, Van Thong in view of Perez fails to teach the system of claim 2, wherein said ranking module is adapted to evaluate speed of speech within the speech recordings and tag at least one of speech recordings and specific portions thereof accordingly (Bakis col 4 lines 41-64).

Bakis teaches multiple output sentences for a given word or phrase, where each output sentence for a given word or phrase reflects a different emotional emphasis and could be selected automatically, or manually as desired, to create a specific emotional effect. For example, the same output sentence for a given word or phrase can be recorded three times, to selectively reflect excitement, sadness or fear. In further variations, the same output sentence for a given word or phrase can be recorded to reflect different accents, dialects, pitch, loudness or rates of speech. Changes in the volume or pitch of speech can be utilized, for example, to indicate a change in the importance of the content of the speech. The variable rate of speech outputs can be used to select a translation that has a best fit with the spoken phrase.

Therefore, it would have been obvious to one of ordinary skill in the art at the time of the invention a ranking module adapted to evaluate speed of speech and tag an according speech portion. Tagging ranked data would allow for recognition of the best choices of speech recordings and allows for book marking or the use of keywords or key terms as a form of indicating the most acceptable data in a summarized manner. Additionally, recognizing various prosodic features such as pitch, intonation, emotion, stress, speech rate, duration, etc. would allow for the omission of less important words

during the ranking of a speech data set, where a tag can be applied more precisely prior to ranking (i.e. finding a phrase with specific pitch level, duration, and emotion).

Re claim 8, Van Thong fails to teach a ranking module (Perez col 2 lines 1-14).

Perez teaches that the "n" top choices refers to a variable number of possibilities, which can be supplied by a speech recognition engine tagged with a probability or in a ranked order of descending probability of being correct. For example, the comparison might require that the top two choices from each engine match before accepting the first as the accepted and recognized text. Alternatively, if the top choice of each do not match, it might be acceptable if one of the engines has a second choice that matches the other engine's top choice.

Additionally Perez teaches probability determined during the decoding can be used as a score for the match between the input speech and the chosen sentence text.

NOTE: For purposes of prior art, recognizing a tag is construed to be both functionally equivalent and equally effective as ranking a set of data that is tagged with a probability, where the probability is tagged with a score. In order to rank a set of data, a probability is assigned, where recognition of tags would be necessary in order to achieve a proper ranking scheme.

However, Van Thong in view of Perez fails to teach the system of claim 2, wherein said ranking module is adapted to evaluate emotive character of speech within

the speech recordings and tag at least one of speech recordings and specific portions thereof accordingly (Bakis col 4 lines 41-64).

Bakis teaches multiple output sentences for a given word or phrase, where each output sentence for a given word or phrase reflects a different emotional emphasis and could be selected automatically, or manually as desired, to create a specific emotional effect. For example, the same output sentence for a given word or phrase can be recorded three times, to selectively reflect excitement, sadness or fear. In further variations, the same output sentence for a given word or phrase can be recorded to reflect different accents, dialects, pitch, loudness or rates of speech. Changes in the volume or pitch of speech can be utilized, for example, to indicate a change in the importance of the content of the speech. The variable rate of speech outputs can be used to select a translation that has a best fit with the spoken phrase.

Therefore, it would have been obvious to one of ordinary skill in the art at the time of the invention a ranking module adapted to evaluate emotive character of speech and tag an according speech portion. Tagging ranked data would allow for recognition of the best choices of speech recordings and allows for book marking or the use of keywords or key terms as a form of indicating the most acceptable data in a summarized manner. Additionally, recognizing various prosodic features such as pitch, intonation, emotion, stress, speech rate, duration, etc. would allow for the omission of less important words during the ranking of a speech data set, where a tag can be applied more precisely prior to ranking (i.e. finding a phrase with specific pitch level, duration, and emotion).

Re claim 11, Van Thong fails to teach a navigation module adapted to rank (Perez col 2 lines 1-14).

NOTE: For purposes of prior art, a navigation module is construed to both functionally equivalent and equally effective as the system in (Van Thong Fig. 1), where a user and system respond in accordance with commands and/or text.

Perez teaches that the "n" top choices refers to a variable number of possibilities, which can be supplied by a speech recognition engine tagged with a probability or in a ranked order of descending probability of being correct. For example, the comparison might require that the top two choices from each engine match before accepting the first as the accepted and recognized text. Alternatively, if the top choice of each do not match, it might be acceptable if one of the engines has a second choice that matches the other engine's top choice.

Additionally Perez teaches probability determined during the decoding can be used as a score for the match between the input speech and the chosen sentence text.

NOTE: For purposes of prior art, recognizing a tag is construed to be both functionally equivalent and equally effective as ranking a set of data that is tagged with a probability, where the probability is tagged with a score. In order to rank a set of data, a probability is assigned, where recognition of tags would be necessary in order to achieve a proper ranking scheme.

However, Van Thong in view of Perez fails to teach the system of claim 9, wherein said navigation module is adapted to rank atleast one of speech recordings and specific portions thereof based on quality of pronunciation of speech therein. (Bakis col 4 lines 41-64).

Bakis teaches multiple output sentences for a given word or phrase, where each output sentence for a given word or phrase reflects a different emotional emphasis and could be selected automatically, or manually as desired, to create a specific emotional effect. For example, the same output sentence for a given word or phrase can be recorded three times, to selectively reflect excitement, sadness or fear. In further variations, the same output sentence for a given word or phrase can be recorded to reflect different accents, dialects, pitch, loudness or rates of speech. Changes in the volume or pitch of speech can be utilized, for example, to indicate a change in the importance of the content of the speech. The variable rate of speech outputs can be used to select a translation that has a best fit with the spoken phrase.

Therefore, it would have been obvious to one of ordinary skill in the art at the time of the invention a navigation module that ranks speech based on quality of pronunciation or prosody. Tagging ranked data would allow for recognition of the best choices of speech recordings and allows for book marking or the use of keywords or key terms as a form of indicating the most acceptable data in a summarized manner. Additionally, recognizing various prosodic features such as accent, dialect, pitch, intonation, emotion, stress, speech rate, duration, etc. would allow for the omission of less important words during the ranking of a speech data set, where a tag can be

applied more precisely prior to ranking (i.e. finding a phrase with specific pitch level, duration, and emotion).

Re claim 14, Van Thong fails to teach a navigation module adapted to rank (Perez col 2 lines 1-14).

NOTE: For purposes of prior art, a navigation module is construed to both functionally equivalent and equally effective as the system in (Van Thong Fig. 1), where a user and system respond in accordance with commands and/or text.

Perez teaches that the "n" top choices refers to a variable number of possibilities, which can be supplied by a speech recognition engine tagged with a probability or in a ranked order of descending probability of being correct. For example, the comparison might require that the top two choices from each engine match before accepting the first as the accepted and recognized text. Alternatively, if the top choice of each do not match, it might be acceptable if one of the engines has a second choice that matches the other engine's top choice.

Additionally Perez teaches probability determined during the decoding can be used as a score for the match between the input speech and the chosen sentence text.

NOTE: For purposes of prior art, recognizing a tag is construed to be both functionally equivalent and equally effective as ranking a set of data that is tagged with a probability, where the probability is tagged with a score. In order to rank a set of data, a probability is assigned, where recognition of tags would be necessary in order to achieve a proper ranking scheme.

However, Van Thong in view of Perez fails to teach the system of claim 9, wherein said navigation module is adapted to rank atleast one of speech recordings and specific portions thereof based on duration thereof (Bakis col 2 lines 16-22).

Bakis teaches an event-measuring mechanism measures the duration of various key events in the source phrase. For example, the speech can be normalized in duration using event duration information and presented to the user. Event duration could be, for example, the overall duration of the input phrase, the duration of the phrase with interword silences omitted, or some other relevant durational features.

Therefore, it would have been obvious to one of ordinary skill in the art at the time of the invention a navigation module that ranks speech. Tagging ranked data would allow for recognition of the best choices of speech recordings and allows for book marking or the use of keywords or key terms as a form of indicating the most acceptable data in a summarized manner. Additionally, recognizing various prosodic features such as accent, dialect, pitch, intonation, emotion, stress, speech rate, duration, etc. would allow for the omission of less important words during the ranking of a speech data set, where a tag can be applied more precisely prior to ranking (i.e. finding a phrase with specific pitch level, duration, and emotion).

Re claim 15, Van Thong teaches said navigation module (Fig. 1) and line-specific portion of another speech recording already assigned to a textual line sequentially (col 7 lines 46-60).

However Van Thong fails to teach the system of claim 9, wherein said navigation module is adapted to rank a line-specific portion of a speech (Perez col 2 lines 1-14).

Perez teaches that the "n" top choices refers to a variable number of possibilities, which can be supplied by a speech recognition engine tagged with a probability or in a ranked order of descending probability of being correct. For example, the comparison might require that the top two choices from each engine match before accepting the first as the accepted and recognized text. Alternatively, if the top choice of each do not match, it might be acceptable if one of the engines has a second choice that matches the other engine's top choice.

However, Van Thong in view of Perez fails to teach a recording based on consistency thereof with at least, one adjacent, line-specific portion of another speech recording already assigned to a textual line sequentially adjacent in the script to a textual line aligned to the line-specific portion of the speech recording (Bakis col 2 lines 16-22).

NOTE: For purposes of prior art, an adjacent ranked portion of speech is construed to both functionally equivalent and equally effective as adjacent prosody, such a pitch and rate of speech.

Bakis teaches multiple output sentences for a given word or phrase, where each output sentence for a given word or phrase reflects a different emotional emphasis and could be selected automatically, or manually as desired, to create a specific emotional effect. For example, the same output sentence for a given word or phrase can be

recorded three times, to selectively reflect excitement, sadness or fear. In further variations, the same output sentence for a given word or phrase can be recorded to reflect different accents, dialects, pitch, loudness or rates of speech. Changes in the volume or pitch of speech can be utilized, for example, to indicate a change in the importance of the content of the speech. The variable rate of speech outputs can be used to select a translation that has a best fit with the spoken phrase.

Therefore, it would have been obvious to one of ordinary skill in the art at the time of the invention ranking an adjacent line specific portion of speech assigned to a textual line sequentially adjacent in a script. Tagging ranked data would allow for recognition of the best choices of speech recordings and allows for book marking or the use of keywords or key terms as a form of indicating the most acceptable data in a summarized manner. Additionally, recognizing various prosodic features such as accent, dialect, pitch, intonation, emotion, stress, speech rate, duration, etc. would allow for the omission of less important words during the ranking of a speech data set, where a tag can be applied more precisely prior to ranking (i.e. finding a phrase with specific pitch level, duration, and emotion).

Re claim 16, Van Thong teaches atleast one of speech recordings and specific portions thereof based on ability of thereof to contribute to solutions rendering a combination recording of a target duration and including a partial accumulation of line-specific portions of the multiple speech recordings (col 3 lines 25-31 & Fig. 1 items 13, 15, and 17).

However, Van Thong fails to teach the system of claim 9, wherein said navigation module is adapted to rank (Perez col 2 lines 1-14).

NOTE: For purposes of prior art, a navigation module is construed to both functionally equivalent and equally effective as the system in (Van Thong Fig. 1), where a user and system respond in accordance with commands and/or text.

Perez teaches that the "n" top choices refers to a variable number of possibilities, which can be supplied by a speech recognition engine tagged with a probability or in a ranked order of descending probability of being correct. For example, the comparison might require that the top two choices from each engine match before accepting the first as the accepted and recognized text. Alternatively, if the top choice of each do not match, it might be acceptable if one of the engines has a second choice that matches the other engine's top choice.

However, Van Thong in view of Perez fails to teach a target duration (Bakis col 2 lines 16-22)

Bakis teaches an event-measuring mechanism measures the duration of various key events in the source phrase. For example, the speech can be normalized in duration using event duration information and presented to the user. Event duration could be, for example, the overall duration of the input phrase, the duration of the phrase with interword silences omitted, or some other relevant durational features.

Therefore, it would have been obvious to one of ordinary skill in the art at the time of the invention a navigation module that ranks speech. Tagging ranked data

would allow for recognition of the best choices of speech recordings and allows for book marking or the use of keywords or key terms as a form of indicating the most acceptable data in a summarized manner. Additionally, recognizing various prosodic features such as accent, dialect, pitch, intonation, emotion, stress, speech rate, duration, etc. would allow for the omission of less important words during the ranking of a speech data set, where a tag can be applied more precisely prior to ranking (i.e. finding a phrase with specific pitch level, duration, and emotion).

Re claim 18, Van Thong fails to teach the system of claim 9, wherein said navigation module is adapted to rank at least one of speech recordings and specific portions thereof (Perez col 2 lines 1-14).

Perez teaches that the "n" top choices refers to a variable number of possibilities, which can be supplied by a speech recognition engine tagged with a probability or in a ranked order of descending probability of being correct. For example, the comparison might require that the top two choices from each engine match before accepting the first as the accepted and recognized text. Alternatively, if the top choice of each do not match, it might be acceptable if one of the engines has a second choice that matches the other engine's top choice.

However, Van Thong in view of Perez fails to teach portions based on emotive character exhibited thereby and a target emotive state recorded with respect to a textual line aligned thereto (Bakis col 4 lines 41-64).

Bakis teaches multiple output sentences for a given word or phrase, where each output sentence for a given word or phrase reflects a different emotional emphasis and could be selected automatically, or manually as desired, to create a specific emotional effect. For example, the same output sentence for a given word or phrase can be recorded three times, to selectively reflect excitement, sadness or fear. In further variations, the same output sentence for a given word or phrase can be recorded to reflect different accents, dialects, pitch, loudness or rates of speech. Changes in the volume or pitch of speech can be utilized, for example, to indicate a change in the importance of the content of the speech. The variable rate of speech outputs can be used to select a translation that has a best fit with the spoken phrase.

Therefore, it would have been obvious to one of ordinary skill in the art at the time of the invention a navigation module that ranks speech based on emotive character and a target emotive state. Tagging ranked data would allow for recognition of the best choices of speech recordings and allows for book marking or the use of keywords or key terms as a form of indicating the most acceptable data in a summarized manner. Additionally, recognizing various prosodic features such as pitch, intonation, emotion, stress, speech rate, duration, etc. would allow for the omission of less important words during the ranking of a speech data set, where a tag can be applied more precisely prior to ranking (i.e. finding a phrase with specific pitch level, duration, and emotion).

**6. Claim 19 rejected under 35 U.S.C. 103(a) as being unpatentable over Van Thong US 6490553 B2 (hereinafter Van Thong) in view of Perez-Mendez et al US**

**5754978 A (hereinafter Perez) and further in view of Goldberg US 6223158 B1  
(hereinafter Goldberg).**

Re claim 19, Van Thong in view of Perez fails to teach the system of claim 9, wherein said navigation module is adapted to rank at least one of speech recordings and specific portions thereof in accordance with user-specified weights respective of multiple ranking criteria (Goldberg col 17 lines 45-64).

Goldberg teaches that after forming this candidate set of identifiers, CPU 40 may rank the members of this set from highest to lowest in terms of their respective associative weightings and then prompt the user with each of these ranked identifiers until either the user positively confirms one of these candidate identifiers as matching the input identifier or the user has been prompted with all the candidate identifiers, in which case CPU 40 would issue through voice prompt device 25 an error message.

Therefore, it would have been obvious to one of ordinary skill in the art at the time of the invention ranking speech recordings with user-specified weights respective of multiple ranking criteria. Ranking and weighting speech recordings relevant to a user would allow for a more customized arrangement of data than if a system ranked and chose matches independently. By utilizing user specified weights in addition to a ranking system, a less ambiguous acquisition of speech will be present, where a user has a say during a ranking process.

**7. Claim 24 rejected under 35 U.S.C. 103(a) as being unpatentable over Van Thong US 6490553 B2 (hereinafter Van Thong) in view of Mercs et al US 5999906 A (hereinafter Mercs).**

Re claim 24, Van Thong teaches the system of claim 1, wherein the combination recording includes at least one Voice track of a multiple track (col 2 lines 32-46) audio visual recording (col 5 lines 13-17), the speech recordings are produced, and each speech recording is automatically temporally aligned to the combination recording (col 5 lines 45-62).

However, Van Thong fails to teach a dubbing process (Mercs col 2 lines 24-39).

Mercs teaches Output audio may be to film, tape, speakers and the like. In typical film dubbing applications, an operator will handle multiple channels of audio, simultaneously, by repetitively and selectively punching-in and punching-out on a studio control panel, for both recording and playback, to obtain the desired mix of audio for recording on the sound track of film or video.

Therefore, it would have been obvious to one of ordinary skill in the art at the time of the invention an audio track of audio visual recording produced in a dubbing process where speech is aligned. Aligning speech with audio in a dubbing process is necessary to perform dubbing, where dubbing involves recording, replacing, and aligning audio visual data. Aligning multiple speech recordings with video data would allow for a more robust output of aligned media data.

**8. Claim 25 rejected under 35 U.S.C. 103(a) as being unpatentable over Van Thong US 6490553 B2 (hereinafter Van Thong) in view of Bakis US 6556972 B1 (hereinafter Bakis).**

Re claim 25, Van Thong teaches system of claim 1, wherein the textual lines are sequentially related and the combination recording includes at least one audio track (col 2 lines 32-46).

However, Van Thong fails to teach having a durational constraint. (Bakis col 7 lines 50-62).

Bakis teaches that the duration of the input phrases or the output phrases, or both, can be adjusted in accordance with the present invention. It is noted that it is generally more desirable to stretch the duration of a phrase than to shorten the duration. Thus, the present invention provides a mechanism for selectively adjusting either the source language phrase or the target language phrase. Bakis also teaches that block 850 determines whether the source language phrase or the target language phrase has the shorter duration, and then increases the duration of the phrase with the shorter duration.

Therefore, it would have been obvious to one of ordinary skill in the art at the time of the invention combination recording with an audio track having a durational constant. Constraining the duration of a phrase or speech recording would allow for proper compression or expansion during the alignment of audio, and/or text, and/or video, to allow for precise time alignment.

**9. Claims 26-27 rejected under 35 U.S.C. 103(a) as being unpatentable over Van Thong US 6490553 B2 (hereinafter Van Thong) in view of Sukkar US 6292778 B1 (hereinafter Sukkar).**

Re claim 26, Van Thong fails to teach the system of claim 1, wherein the combination recording includes a navigable set of voice prompts (Sukkar col 7 line 37 – col 8 line 10).

Sukkar teaches an adjunct processor that prompts the user to identify by voice the type of service requested. The speech recognizer 100 receives the speech information, and recognizes the service request. In response thereto, the adjunct processor may further prompt the user to recite some other information.

Therefore, it would have been obvious to one of ordinary skill in the art at the time of the invention a navigable set of voice prompts. Using voice prompts would allow for a system to be updated and trained with new audio material, such as personal user information if desired.

Re claim 27, Van Thong fails to teach the system of claim 1, wherein the combination recording includes a set of training data (Sukkar col 2 lines 42-63) for at least one of a speech synthesizer and a speech recognizer (Sukkar col 7 lines 25 – 36).

Sukkar teaches an adjunct processor that may have additional equipment (not shown), including peripheral equipment, for performing tasks in addition to speech recognition (e.g., for speech synthesis or announcements), for interfacing to other network equipment, and for providing "housekeeping," operating system, and other

functions of a general-purpose computer. Additionally, Sukkar teaches an ASR system can reliably be applied to many different tasks without the need for retraining. If the ASR system is to be used to recognize speech in a language for which it was not originally trained, it may be necessary to update the language model, but because the number of unique subwords is limited, the amount of training data required is substantially reduced.

Therefore, it would have been obvious to one of ordinary skill in the art at the time of the invention training data for a speech recognizer and synthesizer. Using a speech recognizer and synthesizer relevant to training data would allow for a less ambiguous system that can consistently learn and omit less important words when speech is acquired, and a more robust synthesis of speech during output, where the training data will be updated constantly relevant to the input, where new language portions can be implemented that can reduce the need for retraining.

### ***Conclusion***

10. The prior art made of record and not relied upon is considered pertinent to applicant's disclosure. US 6438522 B1, US 5455889 A, US 20060149558 A1, US 6477491 B1, US 20020059148 A1, US 5918222 A, US 6665640 B1, US 20030229497 A1, US 6903723 B1, US 6192343 B1.

Any inquiry concerning this communication or earlier communications from the examiner should be directed to Michael C. Colucci whose telephone number is (571)-270-1847. The examiner can normally be reached on 9:30 am - 6:00 pm, Monday-Friday.

If attempts to reach the examiner by telephone are unsuccessful, the examiner's supervisor, Richemond Dorvil can be reached on (571)-272-7602. The fax phone number for the organization where this application or proceeding is assigned is 571-273-8300.

Information regarding the status of an application may be obtained from the Patent Application Information Retrieval (PAIR) system. Status information for published applications may be obtained from either Private PAIR or Public PAIR. Status information for unpublished applications is available through Private PAIR only. For more information about the PAIR system, see <http://pair-direct.uspto.gov>. Should you have questions on access to the Private PAIR system, contact the Electronic Business Center (EBC) at 866-217-9197 (toll-free). If you would like assistance from a USPTO Customer Service Representative or access to the automated information system, call 800-786-9199 (IN USA OR CANADA) or 571-272-1000.

Michael Colucci Jr.  
Patent Examiner  
AU 2626  
(571)-270-1847  
Michael.Colucci@uspto.gov



RICHEMOND DORVIL  
SUPERVISORY PATENT EXAMINER